GOO vs GPS

Simón Hincapié

Santiago Estrada Bernal

INFORME DE HALLAZGOS

Semillero Coordinadora

Universidad EIA

Envigado, Antioquia

Ingeniería de Sistemas y Computación

2024

1. Análisis de Datos

Al inicio del reto, para el análisis de GOO y GPS, se nos proporcionaron dos archivos diferentes: ***resultadosGOO*** y ***resultadosGPS***. Ambos contienen información (como puede ser el código de remisión asociado a la entrega, la fecha y hora en que fue realizada, coordenadas de entrega, datos del repartidor y el equipo/terminal a la que pertenece, entre otras) acerca de distintas entregas de paquetes realizadas por la empresa.

Inicialmente, se logra evidenciar algunas entradas que contienen datos fuera de lo común o simplemente ilógicos en el contexto de la información que se está guardando. Se observa también una pequeña cantidad de entradas duplicadas (es decir, distintos registros asociados a un mismo código de remisión), como se puede observar en la siguiente imagen extraída de ***resultadosGPS***:

24	22/1/2024, 12:22:54	00483185291	4.5976533,-74.1522032	datorres@coordinadora.com	200	1
25	22/1/2024, 12:21:22	00483185292	4.5978813,-74.1528962	datorres@coordinadora.com	200	1
26	22/1/2024, 12:36:03	00483185292		datorres@coordinadora.com	200	1
27	22/1/2024, 14:19:16	00483185292	4.6036183,-74.1586773	datorres@coordinadora.com	200	1
28	22/1/2024, 13:30:31	00483185293	4.6969184,-74.1237068	diegoatellez@hotmail.com	1	1

Las entradas resaltadas en amarillo estan asociadas al mismo código de remisión (483185292), y se logra observar lo siguiente:

- Todas las entradas tienen una hora de entrega y coordenadas diferentes
- La segunda entrada no tiene coordenadas asociadas
- Todas pertenecen al mismo repartidor

Y así mismo se pueden encontrar entradas atípicas a lo largo de ambos archivos.

Tras un proceso de limpieza inicial (detallado abajo), se procede a encontrar la distancia de haversine (es decir, la distancia entre la latitud y la longitud asociadas a dos puntos distintos en la superficie de una esfera, en este caso la tierra, teniendo en cuenta la curvatura de ésta) entre las entradas asociadas a un mismo código de remisión en ambos archivos. Al realizar este proceso, se logra evidenciar una gran cantidad de incongruencias entre los datos de ambos archivos, ya que muchas entradas presentaron información considerablemente distinta para un mismo código de remisión, cuando deberían ser la misma.

Un ejemplo de esto se puede ver en la siguiente imagen:

	A	B	C	D	E	F	G
1	Codigo Remision	Terminal	Equipo	Distancia Haversine (km)	Diferencia Tiempo (horas)	Fecha_GOO	Fecha_GPS
2	483185268	1	1	10.93573778	0.886343333	2024-01-22 19:48:11	2024-01-22 18:55:00
3	483185516	3	3	9.250013009	6.743989722	2024-01-22 14:35:57	2024-01-22 07:51:19
4	483186176	29	175	7.791706517	10.90813528	2024-01-22 17:17:55	2024-01-22 06:23:26

Para estas 3 entradas, se logra ver como la distancia haversine (la columna en rojo) es considerablemente alta para todas, más aún considerando que, idealmente, los datos de GPS y de GOO deberían tener las mismas coordenadas.

Una vez observadas estas incongruencias, se procede a realizar un proceso de limpieza más exhaustivo para lograr separar los datos atípicos de los no atípicos, el cual se detalla en el siguiente numeral.

2. Proceso de Limpieza

En primer lugar, para poder encontrar la distancia haversine entre las entradas de GPS y GOO se separaron las entradas que no tenían coordenadas asociadas (es decir, que estaban en blanco) en los archivos ***GOO_sinGeo*** y ***GPS_sinGeo***.

Más adelante se le asoció a estos códigos de remisión sin coordenadas la latitud y longitud encontradas en el archivo del otro geocoder (es decir, para las entradas de ***GOO_sinGeo*** se buscó si el código de remisión asociado a estas tenía coordenadas en el archivo ***resultadosGPS*** y viceversa). Los resultados de esto se guardaron en los archivos ***GOO_sinGeoCorregido*** y ***GPS_sinGeoCorregido***.

Una vez realizada la limpieza de los datos en blanco, se construyó el archivo ***distanciaHaversine***. Este archivo tenía en cuenta todos los datos que sí tenían coordenadas asociadas, incluyendo los duplicados. Cabe resaltar que, para la construcción de este archivo, fue necesario realizar una buena cantidad de ajustes a los datos tomados de ambos archivos, ya que estos son muy distintos en la manera en que presentan la información. A continuación se listan los ajustes más importantes:

- Fueron necesarias conversiones y ajustes en algunas columnas de los archivos para poder operar correctamente con los datos
- Los datos de GOO obtienen la fecha de acuerdo al formato UTC, el cual está 5 horas adelantado al formato usado en el archivo de GPS, que cuenta con la zona horaria de Colombia (GMT-5)
- Las coordenadas de GOO no estaban siendo tomadas correctamente, por lo que se tuvo que operar con estos valores para poder trabajar con los extraídos del archivo de GPS (esto se puede ver en el código proporcionado asociado con este archivo, ***haversine.py***)

La primera limpieza que se le realizó a este archivo fue eliminar estas entradas duplicadas que corresponden a un mismo código de remisión, dejando la que tuviera una menor diferencia de tiempo entre la fecha reportada en GPS y la fecha reportada en GOO. Además, en una revisión realizada a este archivo, se logró evidenciar que las entradas asociadas al empleado coordigoo@gmail.com presentaban datos incoherentes, por lo que se decidió clasificarlas como datos atípicos. Los otros criterios para considerar una entrada como atípica fueron los siguientes:

- Si una entrada tenía una diferencia de tiempo mayor a 5 minutos entre la fecha registrada en GOO y la fecha registrada en GPS, era considerada como atípica
- Si una entrada tenía una velocidad promedio mayor a 80km/h, se consideraba como atípica

Cabe resaltar que ninguno de los archivos contiene datos asociados a la velocidad promedio. Esta fue calculada con la diferencia de tiempo y la distancia haversine obtenidas de los datos de ambos archivos.

En la última versión del archivo (es decir, ***distanciaHaversineSinDuplicados***), se guardan las siguientes columnas:

Codigo Remision	Terminal	Equipo_GPS	Equipo_GOO	Empleado_GPS	Empleado_GOO

Distancia Haversine (m)	Diferencia Tiempo (minutos)	Velocidad (km/h)	Fecha_GOO	Fecha_GPS	Atipico

Es importante resaltar que se encontraron datos que presentaban discrepancias en los equipos/terminales registrados (eran distintos los presentados en GPS a los de GOO), pero dado que muchos de estos presentaban datos coherentes tanto en la distancia haversine como en la diferencia de tiempo, no se consideraron como atípicos.

Hay otra versión de este mismo archivo que contiene las coordenadas de tanto GPS como GOO, llamado ***conCoords***

3. Distancias Entre GOO y GPS

Una vez construidos los archivos, se procedió a encontrar los promedios de las distancias haversine registradas. Estos promedios se calculan para dos casos distintos, uno es teniendo en cuenta datos atípicos y el otro es sin tenerlos en cuenta. Para ambos casos se calculó el promedio para cada terminal. Los promedios obtenidos fueron los siguientes:

Terminal	Promedio Distancia Haversine (m)	conAtipicos
1	66.58307514	NO
2	192.7879678	NO
3	98.80052995	NO
4	164.3239951	NO
6	128.3902097	NO
12	197.0197636	NO
13	111.021632	NO
21	137.3386465	NO
22	1.927023194	NO
29	72.88517358	NO
31	18.48371717	NO
General	124.7322435	NO
1	398.8491537	SI
2	241.4177823	SI
3	135.5818914	SI
4	192.8312203	SI
6	214.8943661	SI
12	281.530285	SI
13	193.3143093	SI
21	250.7409389	SI
22	1.927023194	SI
29	126.3215583	SI
31	136.9841037	SI
General	232.0630181	SI

Se logra evidenciar que, teniendo en cuenta datos atípicos, los promedios se ven alterados considerablemente (ver terminales 1 y 21 por ejemplo). Aún así, los promedios que se encuentran a nivel general presentan valores no muy altos (124m sin atípicos y 232m con atípicos), lo que da a entender que la mayoría de las entradas tienen información coherente.

Es importante resaltar que tras la limpieza inicial en la que se eliminaron datos con coordenadas vacías, la terminal 22 resultó con una única entrada, lo cual explica su promedio tan bajo.

4. GOO vs Otros Geocodificadores

El mayor número de entradas resueltas corresponde a GPS (641), lo que indica una tasa de coincidencia del 100% con GOO. Por el contrario, el menor número de entradas resueltas corresponde a Nominatim (70), lo que sugiere un solapamiento limitado entre este geocodificador y GOO en términos de ID con datos completos de latitud y longitud. Otros geocodificadores tienen tasas de coincidencia variables, con Servi (572), Geoapify (578) y Here (605) con tasas de coincidencia relativamente altas, lo que indica una coincidencia sustancial con GOO.

En cuanto a la distancia media entre puntos coincidentes, GPS (232.063 metros) tiene la distancia media más baja entre puntos en los que coincide con GOO, lo que indica que cuando ambos geocodificadores resuelven el mismo ID, las ubicaciones están muy próximas. Por el contrario, GeocodeEarth (30.275,2 metros) tiene la distancia media más alta, lo que indica discrepancias significativas entre las ubicaciones resueltas por él y por GOO. Otros geocodificadores muestran diversos grados de concordancia, con Servi (326,6 metros) con una distancia media relativamente baja, mientras que Geoapify (28.461,3 metros) y Here (6.341,6 metros) muestran mayores discrepancias.

La distancia mínima entre puntos coincidentes es de 0 metros para GPS y GOO, lo que indica coincidencias exactas para algunas identificaciones. Todos los demás geocodificadores tienen distancias mínimas inferiores a 10 metros, lo que indica que, al menos para algunas identificaciones, los geocodificadores y GOO resuelven casi la misma ubicación. En concreto, Servi (0,8758 metros) tiene una distancia mínima muy pequeña, lo que indica una concordancia muy estrecha para algunas identificaciones.

La distancia máxima es la más alta para Geoapify (698.337,3 metros), lo que sugiere la presencia de valores atípicos significativos en los que las ubicaciones resueltas están muy alejadas. GPS (10.935,7 metros) tiene una distancia máxima relativamente menor, lo que indica menos valores atípicos extremos en comparación con Geoapify. GeocodeEarth (554.157,7 metros) y Here (642.766,6 metros) también muestran distancias máximas significativas, lo que indica algunas grandes discrepancias. Otros geocodificadores, como Nominatim (393.622,9 metros) y Servi (10.998,2 metros), también muestran grandes discrepancias en algunos casos.

GPS es el geocodificador que más se asemeja a GOO, tanto en la tasa de coincidencia como en la proximidad de las ubicaciones resueltas. Esto se evidencia por la alta cantidad de

entradas correctamente resueltas, la menor distancia promedio y una distancia mínima de cero. Por otro lado, Nominatim tiene la menor concordancia con GOO, mostrando la tasa de coincidencia más baja y las distancias promedio y máxima más altas, lo que indica grandes diferencias en los datos resueltos por este geocodificador en comparación con GOO. Servi también muestra una buena correspondencia con GOO, con una alta tasa de coincidencia y una baja distancia promedio, lo que sugiere que frecuentemente resuelve ubicaciones similares a las de GOO. Tanto Geoapify como Here presentan discrepancias significativas con GOO, como lo indican sus altas distancias promedio y máximas, lo que sugiere que, aunque puedan resolver muchas de las mismas ubicaciones, estas pueden diferir considerablemente. GeocodeEarth se encuentra en un punto intermedio, con tasas de coincidencia y distancias promedio moderadas, sin coincidir tan estrechamente como GPS o Servi, pero no tan divergentes como Nominatim.

605	389	641	572	70	578	Resueltos
6341.561746	30275.24786	232.0630181	326.6191435	13067.47286	28461.3917	Promedio
1.471255257	1.63539897	0	0.8758092216	3.783602352	5.00081279	Mínimo
642766.6164	554157.6663	10935.73778	10998.20242	393622.9695	698337.3293	Máximo

*En la tabla, de izquierda a derecha, la información corresponde a los siguientes georreferenciadores: here, GeocodeEarth, GPS, servi, nominatim y geopify.

5. GPS vs Otros Geocodificadores

El mayor número de entradas resueltas corresponde a GOO (641), lo que indica una tasa de coincidencia del 100% con GPS. Por el contrario, el menor número de entradas resueltas corresponde a Nominatim (70), lo que sugiere un solapamiento limitado entre este geocodificador y GPS en términos de ID con datos completos de latitud y longitud. Otros geocodificadores tienen tasas de coincidencia variables, con Servi (572), Geoapify (578) y Here (605) con tasas de coincidencia relativamente altas, lo que indica una coincidencia sustancial con GPS.

En cuanto a la distancia media entre puntos coincidentes, GOO (232,063 metros) tiene la distancia media más baja entre puntos en los que coincide con GPS, lo que indica que cuando ambos geocodificadores resuelven el mismo ID, las ubicaciones están muy próximas. Por el contrario, GeocodeEarth (30.254,76 metros) tiene la distancia media más alta, lo que indica discrepancias significativas entre las ubicaciones resueltas por él y GPS. Otros geocodificadores muestran distintos grados de concordancia, con Servi (165,4850 metros) con una distancia media relativamente baja, mientras que Geoapify (28.456,14 metros) y Here (6.225,01 metros) muestran mayores discrepancias.

La distancia mínima entre puntos coincidentes es de 0 metros para GPS y GOO, lo que indica coincidencias exactas para algunas identificaciones. Todos los demás geocodificadores tienen distancias mínimas inferiores a 10 metros, lo que indica que, al menos para algunas identificaciones, los geocodificadores y GPS resuelven casi la misma ubicación. En concreto, Servi (0,7673 metros) tiene una distancia mínima muy pequeña, lo que indica una concordancia muy estrecha para algunas identificaciones.

La distancia máxima es la más alta para Geoapify (698.371,42 metros), lo que sugiere la presencia de valores atípicos significativos en los que las ubicaciones resueltas están muy alejadas. GOO (10.935,74 metros) tiene una distancia máxima relativamente menor, lo que indica menos valores atípicos extremos en comparación con Geoapify. GeocodeEarth (554.111,34 metros) y Here (642.596,14 metros) también muestran distancias máximas significativas, lo que indica algunas grandes discrepancias. Otros geocodificadores, como Nominatim (404.385,96 metros) y Servi (6.225,22 metros), también muestran grandes discrepancias en algunos casos.

GOO muestra la mayor concordancia con GPS, tanto en términos de tasa de coincidencia como de proximidad de las ubicaciones resueltas. Esto queda patente en la tasa de coincidencia del 100%, la distancia media más baja y la distancia mínima nula. Por otro lado, GeocodeEarth presenta la menor concordancia con GPS, como demuestran la menor tasa de coincidencia y las mayores distancias media y máxima, lo que sugiere diferencias significativas en los datos resueltos por este geocodificador en comparación con GPS. Servi muestra una buena concordancia con GPS, con una tasa de coincidencia alta y una distancia media baja, lo que sugiere que a menudo resuelve ubicaciones cercanas a las de GPS. Tanto Geoapify como Here muestran discrepancias sustanciales con GPS, indicadas por distancias medias y máximas elevadas, lo que sugiere que, aunque pueden resolver muchas de las mismas identificaciones, las ubicaciones pueden ser significativamente diferentes. Nominatim y GeocodeEarth se sitúan en un punto intermedio, con tasas de coincidencia y distancias medias moderadas, no coincidiendo tan estrechamente como GOO o Servi, pero no tan divergentes como GeocodeEarth.

La eliminación de valores atípicos en GOO y GPS contribuye probablemente a su mayor tasa de coincidencia y menor distancia media, lo que pone de relieve la importancia de la limpieza de datos en la precisión de la geocodificación.

605	389	Resueltos	572	70	578	641
6225.006048	30254.75837	Promedio	165.4850847	13122.32186	28456.13434	232.0630181
1.165380437	1.976360529	Mínimo	0.7672621923	2.390209199	1.336466579	0
642596.1356	554111.3353	Máximo	6225.222377	404385.9582	698371.4196	10935.73778

*En la tabla, de izquierda a derecha, la información corresponde a los siguientes georreferenciadores: here, GeocodeEarth, GPS, servi, nominatim, geopify y GOO.